Analysing Political Opinions Using Redescription Mining

Esther Galbrun, Pauli Miettinen

Analysing Political Opinions Using Redescription Mining

Esther Galbrun
Inria Nancy – Grand Est
Nancy, France
esther.galbrun@inria.fr

Pauli Miettinen
Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
pauli.miettinen@mpi-inf.mpg.de

***Abstract*—Understanding the socio-economical background of voters supporting a certain cause or, vice versa, understanding the political stance of people from a certain socio-economical niche are important questions in political sciences. Traditionally, answering these questions has required the researcher to fix either the political stance or the socio-economical background. In this paper, we propose using redescription mining to automatically find the stances and niches that correspond to each other. We show how redescription mining can be applied to open data from voting advice applications, providing insights about the position of the candidates to parliamentary elections. Furthermore, we show that these insights are not only descriptive, but that they also generalize well to new data.**

I. Introduction

Are people from low-income classes more likely to support tax reductions? Is there any feature common to people supporting high levels of development aid? These and similar questions are traditionally answered via opinion polls. Conducting polls is expensive, however, and the polls are usually designed to answer at most a handful of questions. Today we have much more data about people's opinions than any single poll could collect, and utilizing this data is the key to understanding the current political opinions and trends. Such data typically does not constitute an unbiased sample, is noisy, and has many other issues, but these issues are often alleviated by the volume of the data. A more important problem, if less often mentioned, is that *we do not know what questions to ask*.

We know how to answer, say, the example questions above. But what if we cannot identify the polarizing topics a priori? What if we do not know which socio-economical niches are the interesting ones? In this paper, we propose to use *redescription mining* to address that issue. Broadly speaking, a *redescription* consists of two ways to describe the same set of entities. For example, characterizing a set of people based on their socio-economical status on one hand and characterizing roughly the same set of people based on their political opinions, on the other hand, would constitute a redescription. The goal of redescription mining is to automatically find good redescriptions, that is, redescriptions that are both accurate and cover large-enough portions of the data.

We use data from two Finnish voting advice applications. The data contains background information about the candidates to the Finnish parliamentary elections in 2011 and 2015, as well as their answers to a number of policy questions. The candidates to parliamentary elections are hardly a good sample of the general population; however, our purpose is not to study the opinions of the general Finnish population, but to demonstrate the usability of redescription mining in this kind of political analysis.

In the following, we will first briefly present redescription mining in more formal terms (Section II), before giving a short introduction to Finnish politics (Section III-A), and explaining the data (Section III-B). In Section IV, we present an analysis of our results from the data. Again, our main purpose is to show the applicability of redescription mining to political data analysis, though we do also make interesting observations regarding Finnish politics.

II. Redescription Mining

Redescription mining is a descriptive data analysis task which aims at simultaneously finding multiple descriptions of a previously unspecified subset of entities [16]. This is in contrast with other methods like *emerging pattern mining*, *contrast set mining*, or *subgroup discovery* (see [15] for a unifying survey) and with general classification methods, where target subsets of entities are specified via labels.

We consider data that contain entities with two sets of characterizing variables. For instance, in the setting considered here, the entities might be candidates to a political election with variables characterizing their personal profile (age, gender, profession, party affiliation, etc.) on one hand, and their position on a selection of issues (pension indexation, membership in NATO, developing wind power, etc.). Then, the goal of redescription mining is to find a pair of queries, one query for each set of variables, such that both queries describe (almost) the same set of entities. We refer to the two sets of variables as left and right hand side data, and the queries over them, respectively, as left and right hand side queries.

More formally, variables can be either Boolean, categorical, or numerical. Boolean variables can be interpreted as a truth value assignment in a natural way. For categorical and real-valued variables, truth value assignments are induced by relations $[v = c]$ and $[a \leq v \leq b]$, respectively, where c is some category and $[a, b]$ an interval. These truth assignments and their negations constitute *literals* which can be combined using the Boolean operators $\wedge$ (and) and $\vee$ (or) to form *queries*.

Then, a redescription is simply a pair of queries over variables from the two sets. The support (supp) of a query is the subset of entities for which the query holds true. The *accuracy* of a redescription is measured by the *Jaccard coefficient* (J) of the supports of its two queries; p-values indicating how likely it is to observe such an overlap for independent queries can be used to reject uninteresting redescriptions.

Several algorithms for redescription mining have been proposed over the years, based in particular on mining and pairing itemsets [5], on learning classification trees [16], [20] or on building queries greedily [5]. The redescriptions presented in this paper were obtained with tree based algorithms [20] and with the REREMI algorithm [5], since these algorithms are able to handle numerical and categorical data as well as missing values, as found in the datasets of interest here. Tree based algorithms build queries by learning classification and regression trees (CART) over either sets of variables in turn while alternating between the two sides, at each step trying to best match the labels assigned by the tree obtained at the previous one. REREMI, on the other hand, is a greedy algorithm that mines redescriptions by iteratively appending new literals to the current queries, at each step keeping the best candidates for further extension. The analysis was carried out using SIREN,[1] an interface that allows to interactively mine, visualize and edit redescriptions [6].

III. THE DATA

A. Introduction to Finnish Politics

The constitution of Finland [2] declares Finland to be a parliamentary representative democratic republic. The unicameral Parliament of Finland has 200 members and the electoral period is four years [2]. Finland has a multi-party system and typically no party holds the majority of the votes.[2] The 2015 parliament, for example, has representatives of eight different parties.[3] As a consequence, the government is usually formed as a coalition of multiple parties, and even the grand coalitions between socialistic and non-socialistic parties are common. The current cabinet (since the 2015 elections) has ministers from three parties, while the cabinet that started after the 2011 elections had ministers from six parties.

The members of parliament are elected every four years on a direct election. For the purpose of the elections, Finland is divided into electoral districts, and the number of seats assigned to each district is proportional to the population. In the 2011 parliamentary elections, there were 14 districts in the mainland but this was reduced to 12 in the 2015 elections. The electoral districts are shown in Figure 1.

Within each electoral district, the seats are allocated following D'Hondt's method [7]: The votes given for all candidates of a party are summed up and the candidates within the party are ordered based on the number of votes they received.

[1] http://siren.gforge.inria.fr/main/

[2] So far it has only happened once that a single party held the majority.

[3] In addition, the constitution [2] assigns one seat to the representative of Åland; this representative is usually not from any of the mainland parties, but works together with the Swedish People's Party of Finland.

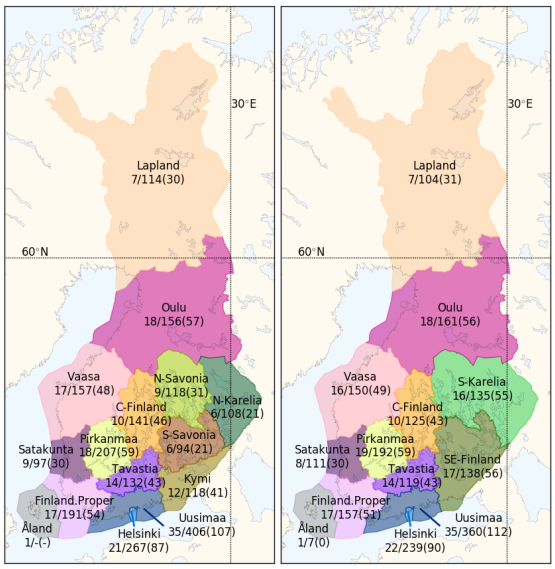


Figure 1. Electoral districts of Finland in 2011 (left) and 2015 (right) parliamentary elections. The numbers $xx/yy(zz)$ below the name show the number of seats (xx), the total number of candidates (yy), and the number of candidates in our data (zz).

They are then assigned a *comparison number* that is the total number of votes the party received divided by the candidate's position within the party. The seats of the electoral district are assigned to the candidates who obtained the highest comparison numbers.

The Finnish election system is thus a complete open list: the order of the candidates within the parties' lists are determined solely based on the votes received by the candidates and there are no official thresholds on the minimum number of votes a candidate needs to receive so as to gain a seat. This means that during the campaign, the candidates try to accomplish two tasks: on one hand, they need to earn as many votes to their party as possible, but on the other hand, they need to earn more votes than the other candidates of their party in order to rank higher on the list. Thus, while the candidates generally follow the platform of their party, they might disagree on some topics or have different priorities in order to distinguish themselves from the other candidates of the party.

In the 2011 and 2015 elections, eight parties won seats in the parliament. The (Finnish) abbreviations used for those parties, their English names, and the number of MPs they had in 2011 and 2015 are shown in Table I. The parties are ordered, from top to bottom, following their seating in the parliament, from left to right, as seen by the Speaker.

B. The Data Sets

For our analysis, we use data from two different Finnish voting advice applications (VAA), one from 2011 and one from 2015. VAAs are online services that collect candidates' answers to various policy questions and allow users to find the candidates whose answers are most closely aligned with theirs. Our data contain background information about the

Table I
PARTIES IN THE FINNISH PARLIAMENT. NUMBER OF MPS IS BASED ON THE SITUATION RIGHT AFTER THE ELECTION. THE PARTIES THAT FORMED THE CABINET AFTER THE ELECTION ARE HIGHLIGHTED IN BOLDFACE.

		Number of MPs	
Abbrev.	English name	2011	2015
Vas	Left Alliance	**14**	12
SDP	Social Democratic Party	**42**	34
Vihr	Green League	**10**	15
PS	Finns Party	39	**38**
Kesk	Centre Party	35	**49**
KD	Christian Democrats	**6**	5
Kok	National Coalition Party	**44**	**37**
RKP	Swedish People's Party[a]	**10**	10

[a]Includes Åland's MP in 2011 and 2015.

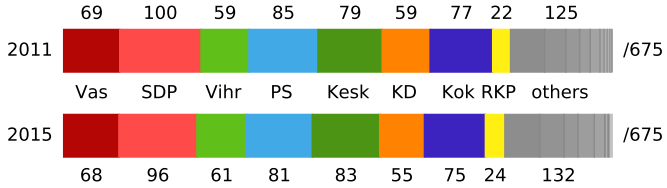


Figure 2. Number of candidates in our data per party in 2011 (HS11 data) and 2015 (Yle15 data) elections. We show the parties that won a seat in the parliament in either of the elections (excluding the representative from Åland); the total number of candidates by the other parties is reported under "others."

candidates (age, party, region, education level, income level, previous political positions, etc.) and their answers to a series of questions. For this work, we considered only the 675 candidates who ran for both the 2011 and the 2015 elections. The distribution of candidates per region is shown in Figure 1, and per party in Figure 2.

The questions in the HS11 data[4] are from the VAA ran by *Helsingin Sanomat* (HS) daily newspaper for the 2011 elections, while the questions in the Yle15 data[5] [19] are from the VAA ran by the Finnish Broadcasting Company *Yle* for the 2015 elections. The HS11 data contains 29 yes–no questions and 2 multiple-choice questions; in addition, the candidates could mark questions which they consider particularly important. The Yle15 data has 86 questions on a 5-point answer scale (disagrees completely–disagrees–neutral–agrees–agrees completely) and 11 yes–no questions.

In addition, we merged the questions of HS11 and Yle15 to form the HS11⋈Yle15 data. That is, in HS11⋈Yle15 we have the answers the candidates gave to the HS11 questions on the left-hand side and their answers to the Yle15 questions on the right-hand side. Queries over the answers in the HS11 data and in the Yle15 data are denoted as **Q11** and **Q15** respectively while queries over the profile information are respectively denoted as **P11** and **P15**. Thus, a redescription from HS11 consists of a pair of queries **P11** and **Q11**, while a redescription from HS11⋈Yle15 consists of a pair of queries **Q11** and **Q15**.

[4]http://blogit.hs.fi/hsnext/hsn-vaalikone-on-nyt-avointa-tietoa, licensed as Creative Commons BY-NC-SA 3.0.

[5]https://www.avoindata.fi/data/fi/dataset/eduskuntavaalien-2015-ylen-vaalikoneen-vastaukset-ja-ehdokkaiden-taustatiedot, licensed as Creative Commons BY 4.0

Table II
SAMPLE REDESCRIPTIONS FROM THE HS11 DATA SET.

Redescription	J	supp
P11: $[24 \leq \text{Age} \leq 58] \vee [7 \leq \text{EduLvl}] \vee \text{RegElected}$ **Q11:** $[\text{Q28.MunNb} \leq 290] \wedge \neg\text{Q31.GvtPrt.KA} \wedge \neg\text{Q31.GvtPrt.PIR} \wedge \neg\text{Q31.GvtPrt.SEN}$	0.833	448
P11: $[51 \leq \text{Age} \leq 58] \vee [7 \leq \text{EduLvl}] \vee \neg\text{MP}$ **Q11:** $\neg\text{Q3.NuclearPow}$	0.656	366
P11: $[\text{Gender} = \text{F}] \vee [1 \leq \text{EduLvl} \leq 5]$ **Q11:** $[\text{Q7.ExtWCareer} = \text{NotAny}]$	0.655	364
P11: $[\text{Party} = \text{Vihr}]$ **Q11:** $\neg\text{Q3.NuclearPow} \wedge [\text{Q10.GreekDebt} = \text{FinInterest}] \wedge [\text{Q20.DvlAid} = \text{Keep}] \wedge \text{Q31.GvtPrt.Vihr}$	0.639	46

IV. ANALYSIS OF THE RESULTS

For the analysis, we used the SIREN software for redescription mining. For all results, we used the RERERMI algorithm; for the policy questions, we allowed only conjunctions over the questions; over the profiles, we allowed arbitrary queries. The data and our full results are freely available.[6]

In the following, we will first present an analysis of some example results from the three data sets. We tested the statistical significance of the redescriptions we found against the null hypothesis that the results are due to a random distribution of the answers following the approach explained in [5]; all of the redescriptions we found were significant at the significance level of 0.01.

The main purpose of these analyses is to demonstrate what kind of insights we can gain by using redescription mining, confirming the usability of redescription mining for the task of political data analysis. We will also use different visualizations of the redescriptions to better interpret them.

A. The 2011 Elections

We present four redescriptions from the HS11 data in Table II.[7] The first redescription has a very high Jaccard coefficient, at 0.833, and relatively high support, of 448 candidates. It says that the candidates who are between 24 and 58 years old, have a university degree, or hold an elected position at the regional level are those who want to limit the municipalities to no more than 290, and do not want three niche parties (*Köyhien Asialla (KA), Piraattipuolue (PIR), Suomen Senioripuolue (SEN)*) in the government. Overall, this redescription covers most of the mainline politicians.

Figure 3(a) shows a decision forest corresponding to this redescription. From it, we can see that the candidates who agree with the questions but do not fit the profile (blue lines) are distributed rather evenly over the part of the profile they

[6]http://people.mpi-inf.mpg.de/~pmiettin/pol_redesc/

[7]The variables in these tables are in abbreviated format, as in the data. The full questions and answers are available in the meta-data of the data sets.

do not fit. On the other hand, the candidates who do fit to the profile, but do not agree with the questions (red lines), mostly disagree with the number of municipalities.

The second redescription concerns the question of whether the parliament should grant permission to build a new nuclear power plant. Here, the question part selects the candidates who are against the power plant, while the profile identifies candidates who are either between 51 and 58 years old, have a university degree, or were not members of the parliament at the time of answering. Despite the very inclusive profile, the redescription still has rather high Jaccard coefficient, at 0.656.

It is interesting to focus on the profiles of those candidates who are against the power plant but do not fit the profile of the redescription. Looking at the blue lines in Figure 3(b), we notice that most of them have a relatively high education level, and most are also above the selected age range.

The third redescription selects candidates who are either female or have lower education levels and who disagree with all the proposed methods for extending work careers, such as increasing the retirement age or removing the option of going on part time retirement. These candidates, in other words, are probably against extending work careers in general. That the concept of extending work careers is less popular among the candidates with lower education (mostly working-class candidates) is not surprising. That female candidates, irrespective of their education, are also against extending work careers is more surprising.

The final redescription in Table II selects the candidates from the Green League. This redescription support consists of 46 candidates out of the 59 Green League candidates. This redescription mostly follows the party platform for the Green League: the party is against the new nuclear power plant, thinks Finland should help Greece deal with the debt crisis, wants to keep the level of development aid stable, and – unsurprisingly – would like to see the Green League in the cabinet. It is perhaps interesting to note that we did not find any other redescription identifying other major parties than the one presented here. This shows the diversity of opinions among the parties' candidates caused by the Finnish election system, as explained in Section III-A.

B. The 2015 Elections

The first redescription in Table III again has a very high Jaccard coefficient, at 0.877. Similarly to the first redescription from the HS11 data, it also describes rather mainline candidates: their election budget is above $1{,}000$ euros, their annual income above $30{,}000$ euros, or they have either no prior political experience or have political experience at the regional level. On the question side, they do not strongly disagree with the request for more funding for the police nor with the claim that Russia is a potential threat to Finland, and do not strongly agree that GMO food is safe.

The second redescription has a simple query over the questions: it selects those candidates who do not strongly agree with having no limits on the opening hours of businesses. The profile, on the other hand, is more complex: it selects only candidates who have a campaign budget lower than $20{,}000$ euros, and among those, the ones who are either from the Communist Party (SKP), are between 47 and 74 years old, or have been elected to a municipal council.

Table III
SAMPLE REDESCRIPTIONS FROM THE YLE15 DATA SET.

Redescription	J	supp
P15: $[1000 \leq$ budged.EUR$] \vee [30000 \leq$ income.EUR$]$ $\vee$pol.exp.none $\vee$ pol.exp.regional.level **Q15:** $[-1 \leq$ Q138.more.police$] \wedge [-1 \leq$ Q140.Russia.is.threat$]$ $\wedge$[Q149.GMO.is.safe ≤ 1]	0.877	449
P15: ([party $=$ SKP$] \vee [47 \leq$ age $\leq 74]$ $\vee$pol.exp.municip.council$) \wedge$ [budget.EUR ≤ 20000] **Q15:** [Q128.no.limit.opening.hours ≤ 1]	0.783	418
P15: $[34 \leq$ age $\leq 74] \vee$ pol.exp.regional.level **Q15:** $[-1 \leq$ Q141.security.more.import.privacy.in.net$]$ $\wedge[1 \leq$ Q147.right.for.healthcare.over.munic.autonomy$]$	0.775	403
P15: $[37 \leq$ age$] \vee$ children **Q15:** [Q143.allow.euthanasia ≤ 1]	0.745	388

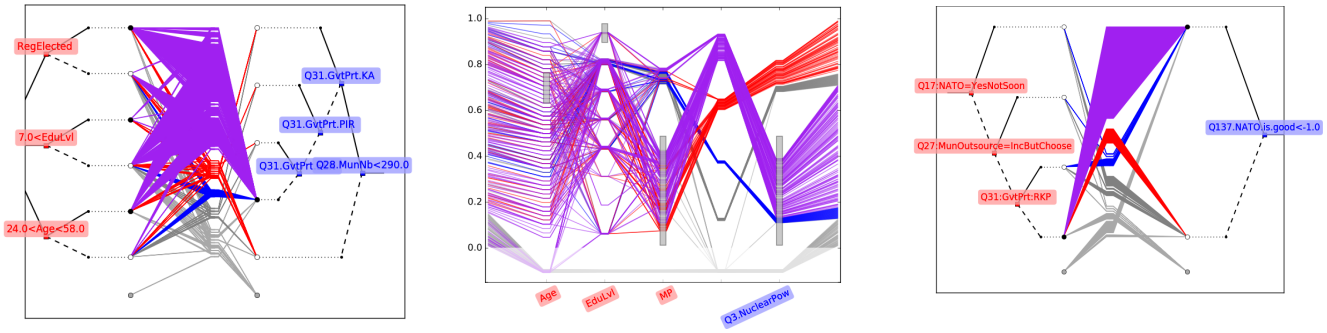
The third redescription again concerns an important policy question in Finland: whether the citizens' right for free, high-quality health care – an important part of the Finnish welfare society – should be enforced by the government, even if that violates the autonomy of the municipalities – an important feature of the Finnish political system. The exact redescription selects the candidates who are between 34 and 74 years old or have experience in regional-level politics and do not strongly disagree with the claim that security is more important than privacy on the Internet (an opinion that is widely shared among candidates in the Yle15 data), and agree that the citizens' right for healthcare is more important than the municipal autonomy.

The final redescription identifies candidates who do not completely agree with allowing euthanasia. Namely, they are the candidates who are either over 37 years old or have children.

C. Joint Data

The HS11$\bowtie$Yle15 data is very different from the previous two data sets. We do not anymore identify candidates based on their profiles and political opinions, but instead based on their answers to questions in 2011 and 2015. This is particularly enlightening, as the questions are sent out by two different media institutions and have a somewhat different emphasis.

The first two redescriptions in Table IV have only one variable on both sides, and they cover popular policy questions. The first redescription selects the candidates who, in 2011, did not think that immigration laws are too strict and in 2015 did not strongly agree that Finland should take more responsibility on immigrants arriving to the EU. That is, these are the candidates who are, if not hostile to immigration, at least reserved towards increasing the number of immigrants. The second redescription selects the candidates who did not want to cut the public spending by reducing the development aid in 2011, and in 2015 did not strongly agree that immigration should be limited because of terrorism threats. These are, in other words, the



(a) Decision forest for the first redescription in Table II.

(b) Parallel coordinates plot for the second redescription in Table II.

(c) Decision forest corresponding to the last redescription in Table IV.

Figure 3. Illustrations of some of the redescriptions. The purple lines correspond to the candidates who agree with both queries (**P11** and **Q11** in HS11 or **Q11** and **Q15** in HS11⋈Yle15), the red lines are candidates who agree with the left-hand query (**P11** in HS11, **Q11** in HS11⋈Yle15), but not the right-hand query, while the blue lines are the candidates who agree with the right-hand query (**Q11** in HS11 and **Q15** in HS11⋈Yle15) but not with the left-hand query.

Table IV
SAMPLE REDESCRIPTIONS FROM THE HS11⋈YLE15 DATA SET.

Redescription	J	supp
Q11: [Q25:Immigration $\neq$ TooStrict] **Q15:** [Q150.more.responsib.EU.immigrants ≤ 1]	0.867	442
Q11: [Q9:CutPblExpend $\neq$ RedDvlAid] **Q15:** [Q139.limit.immigrat.bc.terrorism ≤ 1]	0.848	442
Q11: [Q20:DvlAid $\neq$ Reduce] $\wedge$ $\neg$Q21:Fb:Israel $\wedge\neg$Q31:GvtPrt:KD **Q15:** $\neg$Q246.cancel.gender.neutral.marriage.law	0.803	347
Q11: [Q17:NATO $\neq$ YesNotSoon] $\wedge$ $\neg$Q31:GvtPrt:RKP $\wedge$[Q27:MunOutsource $\neq$ IncButChoose] **Q15:** [Q137.NATO.is.good ≤ -1]	0.776	368

Table V
GENERALIZATION TEST: MINING REDESCRIPTION IN 10-FOLDS. WE REPORT THE NUMBER OF REDESCRIPTIONS MINED (#), THE AVERAGE OF THE RATIO BETWEEN THE ACCURACY (JACCARD COEFFICIENT, J) IN THE TRAINING SET AND THE HOLD-OUT SET, THE AVERAGE ACCURACY OVERALL AND AVERAGE p-VALUE OVERALL ($\pm$ STANDARD DEVIATION).

Dataset	# red.	J ratio	J overall	p-V overall
HS11	62.80 ± 2.52	0.93 ± 0.19	0.58 ± 0.14	0.00 ± 0.00
Yle15	54.70 ± 3.32	0.93 ± 0.15	0.66 ± 0.14	0.00 ± 0.00
HS11⋈Yle15	48.80 ± 4.33	0.95 ± 0.11	0.73 ± 0.05	0.00 ± 0.00

opposite of the above candidates, though many moderates probably fit both of these redescriptions.

The third redescription shows an interesting pattern. In early 2015, the President of Finland signed a new gender-neutral marriage act. Some candidates in the 2015 elections, unhappy that the act passed, proposed to cancel the act. The third redescription shows that the candidates who, in 2015, did not want to cancel the act, are also those who earlier in 2011 did not want to reduce the development aid, did not want Christian Democrats in the government, and did not choose Israel as a country that should be Finland's Facebook friend, if countries would have Facebook accounts. Supporting the gender-neutral marriage act and development aid can be considered as standard liberal political opinions, while the Christian Democrats are typically seen as a rather conservative party, especially when it comes to topics such as marriage.

The fourth redescription concerns a recurring topic in Finnish politics: whether Finland should join the NATO. The redescription selects those candidates who in 2015 strongly disagreed with the claim that joining NATO is good for Finland and who in 2011 did not want to join NATO, did not want to increase the amount of outsourcing the municipalities do, and did not want the Swedish People's Party to participate in the government. The last two opinions are typical of left-wing politicians, although the attitude towards the NATO does not follow the left-right division so clearly. Nonetheless, it is clear from this redescription that the candidates who are against Finland being a member of NATO also have some left-wing policy opinions. The decision forest corresponding to this redescription is presented in Figure 3(c). From it, we see that most of the candidates who in 2015 strongly disagree with the claim that joining the NATO is good for Finland, but did not agree with the claims in 2011 (i.e. the blue lines in Figure 3(c)), would have accepted the Swedish People's Party into the government.

D. Generalization Experiments

So far, we have interpreted our results as *descriptive* patterns, that is, they tell us something about the data as it is. In this section, we test whether our results are also *predictive*, that is, whether they generalize to the opinions of candidates not included in the data. To that end, we performed a test similar to 10-fold cross-validation: we divided the data randomly into 10 equally-sized parts and ran 10 tests. In each test, we held out one part and mined the redescriptions from the training data consisting of the 9 remaining parts. We then computed the Jaccard coefficients of these redescriptions over the testing data, that is, the hold-out part. If the redescriptions have approximately the same Jaccard coefficient in both training and testing data, they are considered to generalize well. For further details about the process, see [20].

The results of the generalization experiments are reported in Table V. They show that in our 10-fold experiment, the redescriptions had very constant quality in both the training and testing data (the ratio of the Jaccard coefficients is almost 1) with relatively high overall Jaccard coefficient. Overall, these results confirm that the redescriptions do not only describe the data, but can also generalize to candidates not covered by the data.

V. Related Work

Voting Advice Applications (VAA) are systems that provide support to citizens in making their voting decision by comparing their position on various political issues to the positions of candidates, whether individual politicians or political parties [12]. The VAAs are often based on asking the users about their opinion on a number of policy statements, and comparing their answers to the positions of the candidates. Beside a list of candidates ranked by proximity, the VAA might display a low dimensional representation of the space of political opinions, for instance as a 2D projection or as radar plots, in which the candidates as well as the users can be placed and compared. The position of the candidates can be obtained externally through expert surveys or manifesto coding – a technique for "normalizing" political programs to allow comparison across time, between parties or between countries. Alternatively, it can be obtained through self-positioning, by having the candidates directly indicate their position regarding the selected statements.

A number of issues pertaining to the construction of VAAs have been pointed out and discussed, such as the selection and wording of the statements, the choice of evaluation, involving both agreement and importance ratings on a fixed scale for instance, the design of the matching algorithm [13] and the choice of the projection space [8], with proposed improvements and alternatives (e.g. [11])

Currently, the number of voting advice applications available online is fairly high, covering a large number of countries and used by millions of users. The Dutch *StemWijzer* [3], created in 1989, is the first known VAA. Finland was the first country to have such a system accessible online, in 1996 [17].

The impact of VAAs on voters' behaviour is also a subject to much debate [10], [18], with multiple studies aiming to analyse the effect of these systems on voters' loyalty to political parties [4], for instance, or evaluate their perceived usefulness [1]. Mendez et al. [14] take a rather practical data analysis point of view on the problem of processing data generated by VAAs to reach sociological conclusions. We note, however, that our approach is different from the approaches considered in these papers, as we do not study the users, but the candidates, and we do not aim at drawing conclusions, e.g., regarding the voting results.

In [9], Grosskreutz et al. employed a data mining method related to the one used here. They applied subgroup discovery to identify relations between socio-economic variables and elections results at the level of voting districts in the german city of Cologne.

VI. Conclusions

In this paper, we have demonstrated the usability of redescription mining for political data analysis. In particular, we have shown how redescription mining can be applied to data from voting advice applications. VAA data, however, is only an example, and we argue that the same methods can be applied to any data with two disjoint sets of variables. In particular, redescription mining can be used with any opinion poll data that records both socio-economical information about the individuals as well as their opinions. Verifying that redescription mining indeed works with other types of data as well is an important topic for future research.

References

[1] R. M. Alvarez, I. Levin, A. H. Trechsel, and K. Vassil. Voting advice applications: How useful and for whom? *J. Inform. Tech. Polit.*, 11(1):82–101, 2014.

[2] Constitution of Finland, translation, 1999. http://www.finlex.fi/en/laki/kaannokset/1999/en19990731.pdf, accessed 3 August 2016.

[3] J. De Graaf. The irresistible rise of stemwijzer. *Voting Advice Applications in Europe: The State of the Art*, pages 35–46, 2010.

[4] Z. Enyedi. The influence of voting advice applications on preferences, loyalties and turnout: An experimental study. *Political Studies*, pages 1467–9248, 2015.

[5] E. Galbrun and P. Miettinen. From Black and White to Full Colour: Extending Redescription Mining Outside the Boolean World. *Stat. Anal. Data Min.*, 5(4):284–303, 2012.

[6] E. Galbrun and P. Miettinen. Siren: An interactive tool for mining and visualizing geospatial redescriptions. In *KDD*, pages 1544–1547, 2012.

[7] M. Gallagher. Proportionality, disproportionality and electoral systems. *Elect. Stud.*, 10(1):33–51, 1991.

[8] M. Germann, F. Mendez, J. Wheatley, and U. Serdült. Spatial maps in voting advice applications: The case for dynamic scale validation. *Acta Politica*, 50(2):214–238, 2015.

[9] H. Grosskreutz, M. Boley, and M. Krause-Traudes. Subgroup Discovery for Election Analysis: A Case Study in Descriptive Data Mining. In *DS '10*, pages 57–71, 2010.

[10] K. Hanel and M. Schultze. Analyzing the political communication patterns of voting advice application users. *Int. J. Internet Sci.*, 9(1):31–51, 2014.

[11] R. Korthals and M. Levels. Multi-attribute compositional voting advice applications (macvaas): a methodology for educating and assisting voters and eliciting their preferences. Technical report, 2016.

[12] S. Marschall and D. Garzia. Voting advice applications in a comparative perspective: an introduction. *Matching Voters with Parties and Candidates.*, page 1, 2014.

[13] F. Mendez. What's behind a matching algorithm: A critical assessment of how vaas produce voting recommendations. *Matching voters with parties and candidates*, 2014.

[14] F. Mendez, K. Gemenis, and C. Djouvas. Methodological challenges in the analysis of voting advice application generated data. In *SMAP '14*, pages 142–148. IEEE, 2014.

[15] P. K. Novak, N. Lavrac, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *J. Mach. Learn. Res.*, 10:377–403, 2009.

[16] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R. F. Helm. Turning CARTwheels: An alternating algorithm for mining redescriptions. In *KDD*, pages 266–275, 2004.

[17] O. Ruusuvirta. Much ado about nothing? Online voting advice applications in Finland. *Voting advice applications in Europe. The state of the art.*, pages 47–77, 2010.

[18] K. Vassil. *Voting Smarter? The impact of voting advice applications on political behavior*. PhD thesis, European University Institute, 2011.

[19] Yle julkaisee vaalikoneen vastaukset avoimena datana, 2015. http://yle.fi/uutiset/yle_julkaisee_vaalikoneen_vastaukset_avoimena_datana/7869597, accessed 12 Aug. 2016.

[20] T. Zinchenko, E. Galbrun, and P. Miettinen. Mining predictive redescriptions with trees. In *ICDMW '15*, 2015.